Research Paper ■

Factors Associated with Success in Searching MEDLINE and Applying Evidence to Answer Clinical Questions

William R. Hersh, MD, M. Katherine Crabtree, RN, DNSc, David H. Hickam, MD, MPH, Lynetta Sacherek, MLS, Charles P. Friedman, PhD, Patricia Tidmarsh, JD, Craig Mosbaek, Dale Kraemer, PhD

Abstract **Objectives:** This study sought to assess the ability of medical and nurse practitioner students to use MEDLINE to obtain evidence for answering clinical questions and to identify factors associated with the successful answering of questions.

Methods: A convenience sample of medical and nurse practitioner students was recruited. After completing instruments measuring demographic variables, computer and searching attitudes and experience, and cognitive traits, the subjects were given a brief orientation to MEDLINE searching and the techniques of evidence-based medicine. The subjects were then given 5 questions (from a pool of 20) to answer in two sessions using the Ovid MEDLINE system and the Oregon Health & Science University library collection. Each question was answered using three possible responses that reflected the quality of the evidence. All actions capable of being logged by the Ovid system were captured. Statistical analysis was performed using a model based on generalized estimating equations. The relevance-based measures of recall and precision were measured by defining end queries and having relevance judgments made by physicians who were not associated with the study.

Results: Forty-five medical and 21 nurse practitioner students provided usable answers to 324 questions. The rate of correctness increased from 32.3 to 51.6 percent for medical students and from 31.7 to 34.7 percent for nurse practitioner students. Ability to answer questions correctly was most strongly associated with correctness of the answer before searching, user experience with MEDLINE features, the evidence-based medicine question type, and the spatial visualization score. The spatial visualization score showed multi-collinearity with student type (medical vs. nurse practitioner). Medical and nurse practitioner students obtained comparable recall and precision, neither of which was associated with correctness of the answer.

Conclusions: Medical and nurse practitioner students in this study were at best moderately successful at answering clinical questions correctly with the assistance of literature searching. The results confirm the importance of evaluating both search ability and the ability to use the resulting information to accomplish a clinical task.

■ **J Am Med Inform Assoc.** 2002;9:283–293.

Affiliations of the authors: Oregon Health & Science University, Portland, Oregon (WRH, MKC, DHH, LS, PT, CM, DK); University of Pittsburgh, Pittsburgh, Pennsylvania (CPF).

This study was supported by grant LM-06311 from the National Library of Medicine.

Correspondence and reprints: William Hersh, MD, Division of Medical Informatics and Outcomes Research, Oregon Health & Science University, BICC, 3181 SW Sam Jackson Park Road, Portland, OR 97201; e-mail: <hersh@ohsu.edu>.

Received for publication: 7/25/01; accepted for publication: 11/29/01.

The MEDLINE database and techniques of evidence-based medicine are increasingly used by health care providers, but little research has elucidated how helpful they are in assisting with clinical decisions. A great deal of work has focused on how well users are able to retrieve relevant documents using information retrieval systems to search MEDLINE, but little work has focused on how well the resulting use of the literature leads to improving ability to answer clinical questions.[1] A number of studies have shown that the techniques of evidence-based medicine can be learned and applied correctly in educational settings,[2] but none has looked at how well they can be applied by students to answer clinical questions.

In the evaluation of information retrieval systems, most studies have focused on measuring the quantities of relevant documents retrieved, using measures of recall and precision. Although useful in measuring retrieval system performance, these measures do not capture the interactive nature of the actual use of systems,[3] tend to focus the assessment on the system and ignore the user,[4] and do not necessarily correlate with user success.[5,6]

A more recent user-centered approach to the evaluation of information retrieval systems has focused on the ability of users to perform tasks with the information retrieval system. The approach assumes that the primary objective of the user is not to retrieve relevant documents but rather to answer questions or obtain new knowledge. The first "task-oriented" evaluation of an information retrieval system was performed by Egan et al.[7] when evaluating the ability of students to answer questions on statistics using the SuperBook hypertext system. Others have subsequently used this general approach to evaluate the abilities of college students to find information in a textbook on Sherlock Holmes[8] and of medical students to answer questions in an online factual database of microbiology.[9,10]

The interactive track at the Text Retrieval Conference (TREC) has adopted a task-oriented framework to assess how well real users can retrieve information from the TREC test collection.[11] This approach has also been used to assess medical students using online textbooks[12] and the MEDLINE database.[13]

The specific research questions addressed in this study were as follows:

- How well are senior medical students and final-year nurse practitioner students able to search MEDLINE with an information retrieval system to answer clinical questions correctly?
- What factors are associated with successful use of an information retrieval system to obtain correct answers to clinical questions?
- Are recall and precision, as measured by conventional recall-precision analyses, associated with successful answering of clinical questions?

Methods

Model

On the basis of results from a prior study,[14] we developed a model of factors that could be associated with the successful answering of questions. Most of these factors were derived from an exhaustive categorization of factors associated with successful use of information retrieval systems, developed by Fidel and Soergel,[15] with some modifications for end-user searching in the health care domain. We also included detailed attributes for determinants of search experience, in particular whether searchers had heard of or used certain advanced MEDLINE features; specifically, Medical Subject Headings (MeSH) terms, subheadings, explosions, and publication types. Table 1 shows the final model of potential predictor factors related to searching ability to be assessed.

The dependent variable in the model is the ability of the user to answer clinical questions correctly. The set of questions for this study was developed in the prior study[14] but modified for conversion to a format that incorporated a judgment of the adequacy of evidence supporting the answer. This was done by wording the questions so they could be answered by one of three statements—"yes, with adequate evidence"; "no, with adequate evidence"; or "insufficient evidence to answer question."

Clinical Questions

The questions used for searching were taken from sources that represented a diverse spectrum of real-world and examination-style information queries. For clinical relevance, the first group of questions was generated by practicing clinicians, and these questions were known to have answers that could be found by searching MEDLINE.[16] We also included some traditional examination-style questions from the Medical Knowledge Self-Assessment Program (MKSAP, American College of Physicians, Philadelphia, Pennsylvania) after converting them from multiple-choice to yes/no form. There were ten questions from each group, which are shown in Table 2.

Table 1 ■

Potential Predicting Factors Influencing Successful Use of an Information Retrieval System by End-users Answering Clinical Questions in a Medical Library Setting Using MEDLINE

Basic::
- ID—user ID
- Question—question number

Answers:
- Answer—answer to question (yes, no, or insufficient evidence to determine)
- Type—EBM type (therapy, diagnosis, harm, prognosis)
- PreAns—answer before searching (yes, no, or insufficient evidence to determine)
- PreCorr—answer correct before search (true, false)
- PreCert—certainty of answer before search (1 = high, 5 = low)
- PostAns—answer after searching (yes, no, or insufficient evidence to determine)
- PostCorr—answer correct after search (true, false)
- PostCert—certainty of answer after search (1 = high, 5 = low)
- Preferred—who user would seek for answer (from list)
- Time—time to complete question (min)
- Stacks—whether searcher went to stacks (true, false)
- Order—order question done by this search (2 to 6; all did same first search, which was ignored)

Questionnaire:
- School—school student enrolled (medical or nurse practitioner)
- Age—(years)
- Sex—(male, female)
- Ethnic—(white or nonwhite)
- CompHrs—computer usage per week (hr)
- ProdSW—use productivity software once a week (yes, no)
- OwnPC—own a personal computer (yes, no)
- Modem—personal computer has a modem (yes, no)
- Internet—personal computer connects to Internet (yes, no)
- LitSrch—literature searches per month (number)
- WebSrch—Web searches per month (number)
- WebMed—Web searches for medical information per month (number)
- TrainEBM—ever had instruction in evidence-based medicine (yes, no)
- HrdMsh—ever heard of MeSH terms (yes, no)

Questionnaire *(cont.):*
- UsedMsh—ever used MeSH terms (yes, no)
- HrdSH—ever heard of subheadings (yes, no)
- UsedSH—ever used subheadings (yes, no)
- HrdExp—ever heard of explosions (yes, no)
- UsedExp—ever used explosions (yes, no)
- HrdPT—ever heard of publication types (yes, no)
- UsedPT—ever used publication types (yes, no)
- PracHard—practice easier or harder with computers (easier, harder)
- EnjComp—enjoy using computers (yes, no)
- MedSpec—medical specialty will be entering (from list, M students only)
- YrsNurse—years worked as a nurse (years, nurse practitioner students only)
- nurse practitionerSpec—nurse practitioner specialty (from list, nurse practitioner only)

Tests:
- VZ2—spatial reasoning test (score)
- RL1—logical reasoning test (score)
- V4—vocabulary test (score)

Articles:
- Helpful—citations helpful answer (number)
- Justified—citations justifying answer (number)

Log:
- Sets—sets in MEDLINE search (number)
- Viewed—total MEDLINE references viewed (number)
- FTViewed—full-text documents viewed (number)

QUIS:
- Quis—QUIS average for this searcher (number)

Retrieval:
- Retrieved—number of articles retrieved by user in terminal set(s)
- Precision—user's precision for retrieval of definitely or possibly relevant articles
- Recall—user's recall for retrieval of definitely or possibly relevant articles

Experimental Protocol

To obtain subjects for the experiment, a convenience sample of senior medical students from Oregon Health & Science University (OHSU) and nurse practitioner students from OHSU and Washington State University–Vancouver was recruited by e-mail, paper mail, and, in the case of nurse practitioner students, announcements in classes. Students were offered remuneration of $100 for successful completion of all tasks.

The general experimental protocol was to participate in three sessions—a "large-group" session where the students would be administered questionnaires and receive an orientation to MEDLINE, the techniques of evidence-based medicine, and the experiment, followed by two hands-on sessions where they would do the actual searching, read the articles, and answer the questions.

The large-group sessions, consisting of 3 to 15 subjects at a time, took place in a computer training room. At each session, subjects were first administered a questionnaire on their personal characteristics and experience with computers and searching factors, from Table 1. Next they were tested for the following cognitive attributes, measured by validated instruments from the Educational Testing Service Kit of Cognitive Factors[17] (ETS mnemonic in parentheses)—paper folding test to assess spatial visualization (VZ-2), nonsense syllogisms test to assess logical reasoning (RL-1), and advanced vocabulary test I to assess verbal reasoning (V-4).

Table 2 ■

Study Questions Grouped by Origin

Questions from clinical practice:

1. Is there any benefit of routine Pap smear in persons who have had a hysterectomy for benign disease?
2. Is ultrasound the best diagnostic test available to exclude the presence of lower extremity deep vein thrombosis?
3. Are nonacetylated salicylates really safer, e.g., have less incidence of acid-peptic problems, in patients with NSAID (nonsteroidal anti-inflammatory drug) gastrointestinal intolerance (who benefit from anti-inflammatory effect)?
4. Is the elevation of alkaline phosphatase a better indicator of recurring prostate cancer than a rising PSA (prostate-specific antigen)?
5. Is the Cytobrush superior to a spatula for obtaining cells for Pap smears, in terms of technical quality (e.g., percentage of interpretable smears)?
6. Does dietary protein effect the level of proteinuria in patients with protein-losing nephropathy?
7. Is there any benefit of ultrasound as physical therapy for sprained ankle?
8. Is penicillin superior to ciprofloxacin for the outpatient treatment of pelvic inflammatory disease?
9. Is anti-inflammatory therapy (NSAIDs) better than Tylenol for elderly patients with degenerative joint disease?
10. Is there evidence of an association between petroleum product exposure and bladder cancer?

Questions derived from medical test questions:

1. Is a high-dose (1,200 to 1,500 mg daily) regimen of zidovudine therapeutically superior to a low-dose (500 to 600 mg daily) one for reducing the progression to AIDs in patients with positive HIV antibody ?
2. Will PSA screening lower the mortality rate from prostate cancer in low-risk men after they reach the age of 50 years?
3. Is there good evidence that an antibiotic can prevent endocarditis in an 18-year-old woman with rheumatic heart disease (mild mitral regurgitation) who is to have a dental root canal?
4. A 52-year-old woman recently had a modified radical mastectomy for infiltrating ductal carcinoma of the breast. Her axillary lymph nodes are negative for tumor. Would estrogen receptor negativity be more likely to indicate a relatively poor prognosis for this patient, rather than thyroid hormone receptor positivity?
5. A 40-year-old premenopausal woman consults you about her risk of breast cancer. Does prior use of birth control pills increase her risk?
6. Does anti-reflux surgery in patients with Barrett's esophagus reduce the risk of developing adenocarcinoma?
7. Is long-distance running associated with intervertebral disc narrowing in men?
8. Would plasma norepinepherine levels indicate poor prognosis in congestive heart failure better than hyponatremia?
9. Is Trental (pentoxifylline) the best drug available to improve symptoms of peripheral vascular disease?
10. Do the majority (> 50 percent) of terminal AIDS patients have clinical symptoms of cardiac involvement?

These cognitive factors were assessed because they have been found to be associated with successful use of computer systems in general and retrieval systems in particular:

- *Spatial visualization*—The ability to visualize spatial relationships among objects has been associated with retrieval system performance by nurses,[18] ability to locate text in a general retrieval system,[19] and ability to use a direct-manipulation (three-dimensional) retrieval system user interface.[20]
- *Logical reasoning*—The ability to reason from premise to conclusion has been shown to improve selectivity in assessing relevant and nonrelevant citations in a retrieval system.[21]
- *Verbal reasoning*—The ability to understand vocabulary has been shown to be associated with the use of a larger number of search expressions and high-frequency search terms in a retrieval systems.[21]

The large-group session also included a brief orientation to the searching task of the experiment as well as a 30-minute hands-on training session covering basic MEDLINE and evidence-based medicine principles. The following searching features were chosen for coverage—MeSH headings, text words, explosions, combinations, limits, and scope notes. These features were chosen because they are taught in medical informatics training courses for health care providers offered at OHSU, and they constitute a basic skill set for MEDLINE searching by a health care provider. The overview of evidence-based medicine described the basic notions of framing the appropriate question, determining which evidence would be most appropriate for a given question, and the best searching strategies for finding such evidence. The teaching was done by a medical informatician experienced in teaching MEDLINE and evidence-based medicine to clinicians (WRH).

The hands-on sessions took place 2 to 4 weeks after the subject had completed the large-group session. He or she had been encouraged to practice the searching skills taught in the large-group session but was given no other explicit instructions. The searching sessions took place in the OHSU Library. All searching was done using the Ovid information retrieval system (Ovid Technologies, New York, New York), which accesses MEDLINE and a collection of 85 full-text journals. We used the Web-based version of Ovid. We also employed its logging facility, which enabled all search statements to be recorded as well as the number of citations presented to and viewed by the user in each set.

In the two hands-on sessions, subjects searched six questions. For the first question of the first session, each user searched the same "practice" question, which was not graded. This was done not only to make searchers comfortable with the experimental process but also because a previous study had suggested a learning effect among inexperienced searchers.[22] The remaining five questions (the last two from the first session and all three from the second session) were selected at random from the pool of 20 questions. Question selection was without replacement, i.e., the same pool of questions was used for four consecutive searchers.

Subjects were limited to one hour per question. Before searching, each subject was asked to record a pre-search answer and a rating of certainty on a scale of 1 (most) to 5 (least) for the questions on which they would search. Subjects were then instructed to perform their searching in MEDLINE and to obtain any articles that they wanted to read either in the library stacks or in the full-text collection available online. They were asked to record on paper their post-search answer, the certainty of their answer (on the 1-to-5 scale), which articles justified their answer, and any article that they looked at in the stacks or in full-text on the screen. On completion of the searching, they were administered the Questionnaire for User Interface Satisfaction (QUIS) 5.0 instrument to measure their satisfaction with the searching system. QUIS measures user satisfaction with a computer system, providing a score from 0 (poor) to 9 (excellent) on a variety of user factors, with the overall score determined by averaging responses to each item.[23]

Searching time for each question was measured using a wall clock. All user–system interactions were logged by the Ovid system software. The search logs were processed to count the number of search cycles (each consisting of the entry of a search term or Boolean combination of sets) and the number of full MEDLINE references viewed on the screen.

Answer Scoring

After all the hands-on searching sessions were completed, the actual answers to the questions were determined by the research team. This was done by assembling all the articles retrieved for each question and giving them, along with the question, to three members of the study team (WRH, MKC, and DHH). The three first designated an answer individually (blinded to any answers that subjects may have provided) and then worked out their differences by consensus. After the answers were designated, two members of the study team (WRH and MKC) graded the answer forms, resolving any differences by consensus. The primary measure of correctness was whether the subjects selected the correct answer from "yes, with adequate evidence"; "no, with adequate evidence"; or "insufficient evidence to answer question."

Statistical Analysis

The usual appropriate statistical analysis for studies with a binary outcome measure (correct vs. incorrect) is logistic regression. However, traditional logistic regression is not appropriate with these data because it does not take into account the within-subject correlation, i.e., the fact that individual questions are not independent, because each searcher answered five questions. To account for this, the analyses were done using generalized estimating equations (GEEs), which account for within-subject correlation.[24] All analyses, both univariate and multivariate, were done using GEE on version 8.01 of the SAS statistical package for Microsoft Windows.

Recall-Precision Analysis

The goal of the recall-precision analysis was to identify a relative measure of recall and precision that could be used to determine its contribution to predicting successful answering of the question. We aimed to carry out the study using the approaches most commonly reported in the information retrieval literature, such as using domain experts to judge relevance, pooling documents within a single query to mask the number of searchers who retrieved it, and assessing interrater reliability. Because of limitations of the retrieval process and of study resources, we were not able to calculate absolute recall and precision. We instead calculated relative measures for each that would allow assessment of their association with successful question answering.

The recall-precision analysis was performed by use of searching logs. The first challenge in this process was to determine which sets to use for each user and question in the analysis. Ovid and other Boolean-oriented systems produce sets of results. Usually, the first sets are large and later ones are smaller, as the search is refined. The user usually does not start looking at the sets until they are smaller and refined. For example, a search on the first question derived from medical test questions shown in Table 2 ("Is a high dose (1,200 to 1,500 mg daily) regimen of zidovudine therapeutically superior to a low dose (500 to 600 mg daily) regimen for reducing the progression to AIDs in patients with positive HIV antibody?") would probably begin with

Table 3 ■

Values of All Searching-related Factors for All Searches, Stratified by Student Type (Medical Student or Nurse Practitioner Student)

Variable	All	Medical	Nurse Practitioner
Number	66	45	21
PreCorr (% true)	32%	32%	31%
PreCert (1–5)	3.2	3.2	3.2
PostCorr (% true)	46%	512%	35%
PostCert (1–5)	2.0	2.0	2.1
Time (min)	32	30	39
Stacks (= Used)	28%	32%	20%
Order (2–6)	4.0	4.0	4.0
School (= M)	68%	100%	0%
Age (yr)	34	31	41
Sex (= M)	35%	51%	0%
CompHrs (hr)	8.4	7.4	10.7
ProdSW (% true)	79%	72%	95%
OwnPC (% true)	92%	91%	95%
Modem (% true)	86%	82%	95%
Internet (% true)	73%	67%	85%
LitSrch (monthly)	5.7	7.1	2.6
WebSrch (monthly)	9.0	11.0	4.5
WebMed (monthly)	2.5	2.4	2.7
TrainEBM (% true)	57%	53%	67%
HrdMsh (% true)	91%	91%	92%
UsedMsh (% true)	73%	75%	67%
HrdSH (% true)	87%	85%	92%
UsedSH (% true)	60%	60%	58%
HrdExp (% true)	79%	83%	72%
UsedExp (% true)	49%	60%	25%
HrdPT (% true)	56%	69%	28%
UsedPT (% true)	22%	29%	5.0%
PracHard (% true)	68%	72%	60%
EnjComp (% true)	88%	87%	90%
VZ2 (score)	12.7	14.2	9.2
RL1 (score)	12.6	14.2	9.0
V4 (score)	22.6	23.1	21.6
Helpful (articles)	2.3	2.3	2.3
Justified (articles)	1.6	1.7	1.4
Sets (number)	19.1	21.6	13.7
Viewed (articles)	8.6	7.1	12.0
FTViewed (articles)	0.94	0.87	1.1
Quis (score)	6.6	6.8	6.3
Retrieved (articles)	26	23	32
Precision (calculated)	29%	30%	26%
Recall (calculated)	18%	18%	20%

sets created with the terms zidovudine and AIDS. Each of these sets yields large numbers of articles, but their combination with AND as well as applications of limits (such as publication type) would yield a more manageable set.

We therefore wanted to restrict our recall-precision calculations to sets that the user would be likely to browse to view specific articles. We thus aimed to identify the "end queries" of the search process, which we identified as the terminal point of a search strategy. This was defined as the point at which the subject stopped refining (creating subsets) of a search and began using new search terms or new combinations of search terms. The document set retrieved by the end queries also had to include the documents cited by the subjects as justification for their post-search answer.

These rules for end queries were given to a graduate medical informatics student who was asked to read the rules and identify end queries in ten systematically selected query sets. The selected query sets represented different users and study questions and were from the beginning, middle, and end of query logs. The graduate student's identification of end queries was compared with the selection of end queries for the same set by a member of the study team (PT). The graduate student and study team member identified 34 end queries. They initially agreed on 23 of the 34 end queries (67.6 percent). The rules were refined by consensus and then applied to all the study logs. End queries that retrieved 200 or more citations were excluded from the relevancy analysis. A total of 10,508 unique question/document pairs were identified and placed in the document pool.

To assess the reliability of the relevance judgment process and determine the number of relevance judges required per question/document pair, a pilot study using 100 documents, selected from a random sampling of five study questions, was performed. The judgments were made by six physicians, all either general internal medicine or medical informatics postdoctoral fellows. All six judges rated the relevance of all 100 documents using a three-point rating scale of "not relevant," "possibly relevant," and "definitely relevant." Using Cronbach's alpha, measured at 0.93, it was determined that three judges per question/document pair were sufficient for reliable assessment of relevance in the larger collection.

To have each question/document pair rated by three judges, we could assess only half (5,254) of the documents retrieved by users, because of limited study

resources. The six judges who participated in the pilot study also participated in the complete study. Three of them judged each unique question/document pair. All judgments were done using the MEDLINE record distributed in an electronic file, although they were encouraged to seek the full text of the article in the library, if necessary.

Results

A total of 66 searchers—45 medical students and 21 nurse practitioner students—performed five searches each, for a total of 330 searches. Six searches were discarded, five because the user did not search MEDLINE and one because the user did not provide an answer, which left 324 searches for analysis.

General Results

There were several differences between medical and nurse practitioner students in this study (Table 3). Use of computers and use of productivity software were higher for nurse practitioner students, but searching experience was higher for medical students. Medical students also had higher self-rating of knowledge and experience with advanced MEDLINE features. Nurse practitioner students tended to be older, and all were female (compared with medical students, of whom 50 percent were female). Medical students also had higher scores on the three cognitive tests. In searching, medical students tended to view more sets but fewer references. They also had a higher level of satisfaction with the information retrieval system, as measured by QUIS.

Prior to searching, the performance of all students was slightly worse than chance, with 104 (32.1 percent) correct and 220 (67.9 percent) incorrect answers. The rate of correctness before searching for medical and nurse practitioner students was virtually identical (32.3 vs. 31.7 percent), as was the rating of certainty (mean, 3.16 for medical students and 3.23 for nurse practitioner students), which was low for both groups.

Following searching, there were 150 (46.3 percent) correct answers and 174 (53.7 percent) incorrect answers. The medical students had a higher rate of correctness than nurse practitioner students (51.6 vs. 34.7 percent). Examination of the results in more detail (Table 4) shows that medical students were better able to use searching to convert incorrect answers into correct ones. Both groups had comparable rates of initially correct answers staying correct or becoming incorrect after searching.

Table 4 ■

Cross-tabulation of Number and Percentage of Incorrect and Correct Answers Before and After Searching, for All Students, Medical Students, and Nurse Practitioner (NP) Students

Pre-search	Post-search	
	Incorrect	Correct
Incorrect:		
All students	133 (41%)	87 (27%)
Medical students	81 (36%)	70 (31%)
NP students	52 (52%)	17 (17%)
Correct		
All students	41 (13%)	63 (19%)
Medical students	27 (12%)	45 (20%)
NP students	14 (14%)	18 (18%)

NOTE: Percentages represent correct answers within each group of students.

Statistical Analysis

The goal of the statistical analysis was to build a model of the factors associated with successful searching, as defined by the outcome variable of correct answer after searching (PostCorr). A GEE model was built after individual variables were screened for their p values, using ANOVA for continuous variables and chi-square tests for categorical variables (Table 5). We also made one adjustment in the data, which was to combine the measures of MEDLINE experience (asking subjects if they had heard of or used four advanced MEDLINE search features—MeSH terms, subheadings, explosions, and publication types) into a set of scale variables. The most statistically predictive scale variable was Used2, which allocated one point if the subject said they had used publication types and one point if they had used explosions in prior MEDLINE searching.

A backward variable selection scheme was performed to determine the best model that predicted correct answering of the question after the MEDLINE search. All variables that predicted the outcome with a p value less than 0.25 were included in the initial backward regression model. The variable with the highest p value was deleted from the model, and the model was then re-run until all variables had p values less than 0.05.

After the backward scheme, variables were put back into the model to see whether any were significant. None of the excluded variables, when added to the

Table 5 ■

Values of Searching-related Factors Stratified by Correctness of Answer, along with *p* Values of Screening, for Statistical Analysis

Variable	Incorrect	Correct	Screening *p* Value
Number	174	150	N/A
PreCorr (% true)	24%	42%	0.00
PreCert (1–5)	3.1	3.3	0.16
PostCert (1–5)	2.0	2.0	0.66
Time (min)	33	32	0.41
Stacks (= Used)	24%	33%	0.10
Order (2–6)	4.0	4.0	0.95
School (= M)	62%	77%	0.01
Age (yr)	34	33	0.55
Sex (= M)	29%	42%	0.03
CompHrs (hr)	8.0	8.8	0.48
ProdSW (% true)	79%	79%	0.95
OwnPC (% true)	93%	91%	0.62
Modem (% true)	87%	85%	0.74
Internet (% true)	73%	72%	0.84
LitSrch (monthly)	4.9	6.7	0.16
WebSrch (monthly)	7.7	10.5	0.05
WebMed (monthly)	2.4	2.7	0.46
TrainEBM (% true)	59%	56%	0.69
HrdMsh (% true)	90%	93%	0.10
UsedMsh (% true)	70%	77%	0.14
HrdSH (% true)	85%	89%	0.18
UsedSH (% true)	56%	64%	0.16
HrdExp (% true)	75%	84%	0.02
UsedExp (% true)	42%	57%	0.01
HrdPT (% true)	49%	65%	0.01
UsedPT (% true)	16%	28%	0.02
PracHard (% true)	68%	69%	0.92
EnjComp (% true)	87%	88%	0.88
VZ2 (score)	12.0	13.4	0.00
RL1 (score)	11.9	13.4	0.10
V4 (score)	22.2	23.1	0.21
Helpful (articles)	2.4	2.2	0.23
Justified (articles)	1.6	1.6	0.57
Sets (number)	21	17	0.29
Viewed (articles)	9.1	8.0	0.14
FTViewed (articles)	0.9	1.0	0.52
Quis (score)	6.6	6.7	0.78
Retrieved (articles)	27	25	0.74
Precision (calculated)	28%	29%	0.99
Recall (calculated)	18%	18%	0.61

final model, had a *p* value less than 0.10. Interaction terms were tested with the final model, and none were significant.

A forward variable selection scheme yielded the same best model. The final model showed that PreCorr, VZ2, Used2, and Type were significant (Table 6). For the variable Type (evidence-based medicine question type), questions of prognosis had the highest likelihood of being answered correctly, followed by questions of therapy, diagnosis, and harm. The analysis also found that the VZ2 and School variables demonstrated multi-collinearity, i.e., they were very highly correlated, and once one was in the model, the other did not provide any additional statistical significance. The VZ2 variable was included in the final model because it led to a higher overall *p* value for the model than School.

Next, a similar analysis was done to find the best model using the subset of cases ($n = 220$) in which the subject did not have the right answer before the MEDLINE search. As shown in Table 7, the final best model was very similar to the model for all questions, with PreCorr obviously excluded. VZ2 and School again showed high multi-collinearity.

To further assess the finding that success in answering questions varied on the basis of evidence-based medicine question type, we looked at the rate of correctness for the four types (Table 8). Because of the exploratory nature of this analysis, we did not perform any statistical analysis. We did find, however, that all subjects did best with prognosis questions, intermediately well with therapy questions, and worst with diagnosis and

Table 6 ■

Statistical Model for All Questions, Including Regression Value with Its *p* Value, Odds Ratio, and 95% Confidence Interval (CI) for the Odds Ratio

Variable	Regression Estimate	*p* Value	Odds Ratio	95% CI
Intercept	−1.71	<0.0001	–	–
PreCorr (correct)	1.07	<0.0001	2.90	1.75–4.80
VZ2	0.0792	0.0136	1.08	1.02–1.15
Used2	0.487	0.0030	1.63	1.18–2.24
Type D	−0.612	0.0794	0.542	0.274–1.07
Type H	−0.613	0.0444	0.542	0.298–0.985
Type P	0.896	0.0140	2.45	1.20–5.01
Type T	0.0000	Overall *p* value for type: <0.0001		

harm questions. The largest gap between medical and nurse practitioner students was with harm and therapy questions; nurse practitioner students did slightly better with diagnosis questions.

Recall–Precision Analysis

Three relevance judgments were made on each of the 5,254 question/document pairs using the responses "not relevant," "possibly relevant," and "definitely relevant." The judges achieved 100 percent agreement (all three judges choose the same rating) for 4,265 of the judgments (81.2 percent), with partial agreement (two of three judges choosing the same rating) for 918 judgments (17.5 percent), and complete disagreement for 71 (1.3 percent).

There were 20 unique groupings of the six relevance judges. For these 20 subsets of relevance judgments, the range of reliability, measured by Cronbach's alpha, was 0.69 to 0.86. The weighted average of these measures was 0.81. Final document relevance was assigned according to the following rules: 1) If all judges agreed, the document was assigned that rating. 2) If two judges agreed, the document was assigned that rating. 3) If all three judges disagreed, the document was assigned the "possibly relevant" rating. The relevance judgments were then used to calculate recall and precision for each user/question pair.

For the 20 questions, 131 documents were judged definitely relevant (average, 6.6 per question) and 528 were judged possibly relevant (average, 26.4 per question). The calculation of recall and precision was done defining relevant documents as those rated definitely or possibly relevant. (Limiting relevance to those defined as definitely relevant only would have left many students' questions with a recall of 0 percent.) As shown in Table 3, there was virtually no difference in recall and precision between medical and nurse practitioner students. Likewise, Table 5 shows that there was no difference in recall and precision between questions that were answered correctly and incorrectly.

Table 7 ■

Statistical Model for Questions Whose Answers Were Incorrect before Searching, Including Regression Value with Its *p* Value, Odds Ratio, and 95% ConfidenceIinterval (CI) for the Odds Ratio

Variable	Regression Estimate	*p* Value	Odds Ratio	95% CI
Intercept	−1.63	0.0026	–	–
VZ2	0.0733	0.0769	1.076	0.992–1.167
Used2	0.436	0.0411	1.546	1.02–2.35
Type D	−0.712	0.104	0.491	0.208–1.16
Type H	−0.304	0.391	0.738	0.369–1.48
Type P	0.701	0.0953	2.015	0.885–4.59
Type T	0.0000	Overall *p* value for type: 0.0151		

Table 8 ■

Rate of Correctness by Evidence-based Medicine Question Type

Question Type	No.	% Correct		
		All Students	Medical Students	NP Students
Diagnosis	60	37	35	41
Harm	64	39	44	25
Prognosis	34	65	70	55
Therapy	166	49	58	38

ABBREVIATION: NP indicates nurse practitioner.

Discussion

This study assessed the ability of a convenience sample of medical and nurse practitioner students to answer clinical questions by searching the literature and using the techniques of evidence-based medicine. We found that this task was challenging for students at this level of experience. They spent an average of more than 30 minutes conducting literature searches and were successful at correctly answering questions less than half the time.

One of the main findings of the study was that medical students were able to use the information retrieval system to improve question answering, while nurse practitioner students were led astray by the system as often as they were helped by it. Another main finding was that experience in searching MEDLINE and spatial visualization ability were associated with the successful answering of questions.

Subjects were also better able to answer certain types of questions in the evidence-based medicine framework than others, doing best with questions of prognosis and worst with those of diagnosis and harm. Another major finding was that the often-studied measures of recall and precision were virtually identical between medical and nurse practitioner stu-

dents and had no association with the correct answering of questions.

Our results showed some similarities to and some differences from a prior study.[14] Somewhat similarly to this study, the prior study found that the most predictive factor of successful question answering was student type (medical vs. nurse practitioner). In that study, spatial visualization showed a trend toward predicting successful answering, but it was short of statistical significance.

In the previous study, unlike this one, the question-answering abilities of both medical and nurse practitioner students improved with use of the information retrieval system. Literature searching experience in that study, as in this one, was associated with the correct answering of questions. Factors that did not predict success in the previous study included age, gender, general computer experience, attitudes toward computers, other cognitive factors (logical reasoning, verbal reasoning, and associational fluency), Meyer-Briggs personality type, and user satisfaction with the information retrieval system. One limitation of the prior study was that it did not assess the application of evidence-based medicine principles in the answering of clinical questions.

The findings of this study are consistent with (and build on) the results of other studies of searching by medical students. Previous studies have shown that training and experience with MEDLINE lead to improved retrieval of relevant articles[25,26] and increased use of MEDLINE in clinical settings.[27] Other than our previous study, described above, there are no other studies of searching by nurse practitioner students.

This study supports the observation that the traditional measures to evaluate information retrieval systems, recall and precision, may have little value in the assessment of how well a system can be used in a real-world setting. While users obviously need to retrieve relevant articles to answer questions, the quantity of relevant articles retrieved had no bearing on the ability to answer them correctly in this study. These findings give credence to those who argue that researchers put too much emphasis on these measures as primary indicators of system efficacy.3–5 They also verify the nonmedical TREC studies that show the same results.[6]

Our results have significant implications for the use of information retrieval systems in clinical settings. The ability to answer clinical questions with the aid of MEDLINE is low. Further research is needed to determine whether additional training, either through the curricula or as part of the study, would change this outcome. Because we used a convenience sample, further research is needed to see whether our findings of differences between medical and nurse practitioner students are generalizable.

For both groups of students, the amount of time taken to answer questions is longer than the amount of time usually devoted to a single patient. Clearly this type of information seeking is practical only "after hours" and not in the clinical setting. Indeed, a growing trend in the evidence-based medicine movement is toward the development of "synthesized" evidence-based content.[28] It may well be that further emphasis should be put on the development of these sorts of information resources for the clinical setting.

One finding of the study, with uncertain meaning, was the strong association of spatial visualization ability with the ability to use an information retrieval system to successfully answer clinical questions. As this variable had such strong multi-collinearity with whether a subject was enrolled in medical or nurse practitioner school, determining which was causal cannot be ascertained from our data.

It may be instructive to explore other results that link computer tasks to spatial visualization. Egan and Gomez[29] have shown that spatial visualization is associated with two processes in text editing—finding the location of characters to be edited and generating a syntactically correct sequence of actions to complete the task. Similarly, Vincente et al.[30] have found that the ability to use a hierarchic file system is associated with spatial visualization as well as with vocabulary skills. In addition, Allen[21] has shown that this trait is associated with the appropriate selection of keywords in searching.

This study had some additional limitations. The use of students, albeit in late stages of their training, limits the generalizability of the results beyond those at their level of clinical training. In future studies, community practitioners will also be included. This study was also limited by taking place in a laboratory setting, in that behaviors in the pursuit of actual clinical knowledge in a real clinical setting may be different from those shown in this controlled environment. However, the ability to use a defined set of tasks and questions provides a benefit that cannot be obtained in the real clinical setting.

In conclusion, this study shows that students in clinical training are at best moderately successful at answering clinical questions correctly with the assis-

tance of searching the literature. Determining the reasons for the limited success of question answering in this study requires further research. The possibilities include everything from inadequate training to an inappropriate database (i.e., a large bibliographic database instead of more concise, synthesized references), problems with the retrieval system, and difficulties in judging evidence. Further studies must develop a priori hypotheses to determine the optimal use of information retrieval systems by clinicians.

References ■

1. Hersh WR, Hickam DH. How well do physicians use electronic information retrieval systems? A framework for investigation and review of the literature. JAMA. 1998;280:1347–52.
2. Norman GR, Shannon SI. Effectiveness of instruction in critical appraisal (evidence-based medicine): a critical appraisal. CMAJ. 1998;158:177–81.
3. Swanson DR. Historical note: Information retrieval and the future of an illusion. J Am Soc Inf Sci. 1988;39:92–8.
4. Harter SP. Psychological relevance and information science. J Am Soc Inf Sci. 1992;43:602–15.
5. Hersh WR. Relevance and retrieval evaluation: perspectives from medicine. J Am Soc Inf Sci. 1994;45:201–6.
6. Hersh W, Turpin A, Price S, et al. Challenging conventional assumptions of automated information retrieval with real users: Boolean searching and batch retrieval evaluations. Inf Proc Manag. 2001;37:383–402.
7. Egan DE, Remde JR, Gomez JM, Landauer TK, Eberhardt J, Lochbaum CC. Formative design-evaluation of Superbook. ACM Trans Inf Syst. 1989;7:30–57.
8. Mynatt BT, Leventhal LM, Instone K, Farhat J, Rohlman DS. Hypertext or book: Which is better for answering questions? Proceedings of Computer-Human Interface '92. 1992:19–25.
9. Wildemuth BM, de Bliek R, Friedman CP, File DD. Medical students' personal knowledge, searching proficiency, and database use in problem solving. J Am Soc Inf Sci. 1995;46:590–607.
10. Friedman CP, Wildemuth BM, Muriuki M, et al. A comparison of hypertext and Boolean access to biomedical information. Proc AMIA Annual Fall Symp. 1996:2–6.
11. Hersh WR. Interactivity at the Text Retrieval Conference (TREC). Inf Proc Manag. 2001;37:365–6.
12. Hersh WR, Molnar A. Toward new measures of information retrieval evaluation. Proceedings of the 18th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval; Seattle, Washington. New York: ACM Press, 1995:164–70.
13. Hersh WR, Pentecost J, Hickam DH. A task-oriented approach to information retrieval evaluation. J Am Soc Inf Sci. 1996;47:50–6.
14. Hersh WR, Crabtree MK, Hickam DH, Sacherek L, Rose L, Friedman CP. Factors associated with successful answering of clinical questions using an information retrieval system. Bull Med Libr Assoc. 2000;88:323–31.
15. Fidel R, Soergel D. Factors affecting online bibliographic retrieval: a conceptual framework for research. J Am Soc Inf Sci. 1983;34:163–80.
16. Gorman PN, Helfand M. Information seeking in primary care: how physicians choose which clinical questions to pursue and which to leave unanswered. Med Decis Making. 1995;15:113–9.
17. Ekstrom RB, French JW, Harmon HH. Manual for Kit of Factor-referenced Cognitive Tests. Princeton, NJ: Educational Testing Service, 1976.
18. Staggers N, Mills ME. Nurse–computer interaction: staff performance outcomes. Nurs Res. 1994;43:144–50.
19. Gomez LM, Egan D, Bowers C. Learning to use a text editor: some learner characteristics that predict success. Human–Computer Interaction. 1986;2:1–23.
20. Swan RC, Allan J. Aspect windows, 3-D visualization, and indirect comparisons of information retrieval systems. Proceedings of the 21st Annual International ACM Special Interest Group in Information Retrieval; Melbourne, Australia. New York: ACM Press, 1998:173–81.
21. Allen BL. Cognitive differences in end-user searching of a CD-ROM index. Proceedings of the 15th Annual International ACM Special Interest Group in Information Retrieval; Copenhagen, Denmark. Copenhagen, Denmark. New York: ACM Press, 1992:298–309.
22. Rose L, Crabtree K, Hersh W. Factors influencing successful use of information retrieval systems by nurse practitioner students. Proc AMIA Annu Fall Symp. 1998:1067.
23. Chin JP, Diehl VA, Norman KL. Development of an instrument measuring user satisfaction of the human–computer interface. Proceedings of CHI '88—Human Factors in Computing Systems. New York: ACM Press, 1988:213–8.
24. Hu FB, Goldberg J, Hedeker D, Flay BR, Pentz MA. Comparison of population-averaged and subject-specific approaches for analyzing repeated binary outcomes. Am J Epidemiol. 1998;147:694–703.
25. Pao ML, Grefsheim SF, Barclay ML, Woolliscroft JO, Shipman BL, McQuillan M. Effect of search experience on sustained MEDLINE usage by students. Acad Med. 1994;69:914–20.
26. Mitchell JA, Johnson ED, Hewett JE, Proud VK. Medical students using Grateful Med: analysis of failed searches and a six-month follow-up study. Comput Biomed Res. 1992;25:43–55.
27. Pao ML, Grefsheim SF, Barclay ML, Woolliscroft JO, McQuillan M, Shipman BL. Factors affecting students' use of MEDLINE. Comput Biomed Res. 1993;26:541–55.
28. Hersh WR. "A world of knowledge at your fingertips": the promise, reality, and future directions of online information retrieval. Acad Med. 1999;74:240–3.
29. Egan DE, Gomez LM. Assaying, isolating, and accommodating individual differences in learning a complex skill. In: Dillon R (ed). Individual Differences in Cognition, vol 2. New York: Academic Press, 1985.
30. Vincente KJ, Leske JS, Williges RC. Assaying and isolating individual differences in searching a hierarchical file system. Human Factors. 1987;29:349–59.